UVA CS 4501: Machine Learning

Lecture 18: Generative Bayes Classifiers

Dr. Yanjun Qi

University of Virginia

Department of Computer Science
Where are we?

Major sections of this course

- Regression (supervised)
- Classification (supervised)
- Unsupervised models
- Learning theory
Where are we? ➡️

Three major sections for classification

• We can divide the large variety of classification approaches into *roughly three major types*

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., *support vector machine*, decision tree

2. Generative:
   - build a generative statistical model
   - e.g., *naïve bayes classifier*, *Bayesian networks*

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. *K nearest neighbors*
Today: Generative Bayes Classifiers

✓ Bayes Classifier
  ▪ MAP classification rule
  ▪ Generative Bayes Classifier
✓ Naïve Bayes Classifier
Review: Notations

• Inputs
  – $X, X_j$ (jth element of vector $X$) : random variables written in capital letter
  – $p$ #input features, $n$ #observations
  – $X$ : matrix written in bold capital
  – Vectors are assumed to be column vectors

• Outputs
  – quantitative $Y$
  – qualitative $C$ (for categorical)
• Treat each feature attribute and the class label as random variables.

• Given a sample \( x \) with attributes \( (x_1, x_2, \ldots, x_p) \):
  – Goal is to predict its class \( c \).
  – Specifically, we want to find the class that maximizes \( p(c \mid x_1, x_2, \ldots, x_p) \).
Review: Bayes Classifiers – MAP Rule

Task: Classify a new instance $X$ based on a tuple of attribute values $X = \langle X_1, X_2, \ldots, X_p \rangle$ into one of the classes $c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid x_1, x_2, \ldots, x_p)$

MAP = Maximum A posteriori Probability

Please read the L13-Logistic for details

Adapt From Carols’ prob tutorial
Review: Establishing a probabilistic model for classification

- (1) Discriminative model

\[ \arg \max_{c \in C} P(c \mid X), \quad C = \{c_1, \ldots, c_L\} \]

Discriminative
Probabilistic Classifier

\[ x = (x_1, x_2, \ldots, x_p) \]
Bayes classifiers

MAP classification rule

- Establishing a probabilistic model for classification
  - (1) Discriminative
  - (2) Generative
A Dataset for classification

\[ f : \{X\} \rightarrow \{C\} \]

\[ \text{Output as Discrete Class Label} \]
\[ C_1, C_2, \ldots, C_L \]

\[ \text{arg max}_{c \in C} P(c \mid X) = \{c_1, \ldots, c_L\} \]

\[ \text{arg max}_{c \in C} P(c \mid X) = \text{arg max}_{c \in C} P(X, c) = \text{arg max}_{c \in C} P(X \mid c) P(c) \]

\[ \text{this lecture!} \]

- Data/points/instances/examples/samples/records: [rows]
- Features/attributes/dimensions/independent variables/covariates/predictors/regressors: [columns, except the last]
- Target/outcome/response/label/dependent variable: special column to be predicted [last column]
Today: Generative Bayes Classifiers

✔ Bayes Classifier
  - MAP classification rule
  - Generative Bayes Classifier
✔ Naïve Bayes Classifier
(2) Generative

\[ P(X|C), \quad C = c_1, \cdots, c_L, X = (X_1, \cdots, X_p) \]

Generative Probabilistic Model for Class 1

\[ P(x|c_1) \]

\[ x_1 \quad x_2 \quad \cdots \quad x_p \]

\[ x = (x_1, x_2, \cdots, x_p) \]

Generative Probabilistic Model for Class 2

\[ P(x|c_2) \]

\[ x_1 \quad x_2 \quad \cdots \quad x_p \]

\[ x = (x_1, x_2, \cdots, x_p) \]

Generative Probabilistic Model for Class L

\[ P(x|c_L) \]

\[ x_1 \quad x_2 \quad \cdots \quad x_p \]

\[ x = (x_1, x_2, \cdots, x_p) \]
Generative BC

\[
P(X|C), \quad C = c_1, \ldots, c_L, \quad X = (X_1, \ldots, X_p)
\]

Generative Probabilistic Model for Class 1

\[
P(x|c_1)
\]

Generative Probabilistic Model for Class 2

\[
P(x|c_2)
\]

Generative Probabilistic Model for Class \(L\)

\[
P(x|c_L)
\]

\[
x = (x_1, x_2, \ldots, x_p)
\]
Generative BC

\[ P(X|C_1), P(X|C_2), \ldots, P(X|C_L) \]

\[ P(X|C), \]
\[ C = c_1, \ldots, c_L, X = (X_1, \ldots, X_p) \]

\[ \Rightarrow p(c|x) \]

MAP rule

Generative Probabilistic Model for Class 1

Generative Probabilistic Model for Class 2

Generative Probabilistic Model for Class L

\[ x = (x_1, x_2, \ldots, x_p) \]
Review: Bayes’ Rule
– for Generative Bayes Classifiers

\[ P(C, X) = P(C \mid X)P(X) = P(X \mid C)P(C) \]

\[ P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \]
Review Probability:
If hard to directly estimate from data, most likely we can estimate

• 1. Joint probability
  – Use Chain Rule

• 2. Marginal probability
  – Use the total law of probability

• 3. Conditional probability
  – Use the Bayes Rule
Review: Bayes’ Rule
– for Generative Bayes Classifiers

\[ P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \]

\[ P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \]

\[ P(C_1 \mid x), P(C_2 \mid x), \ldots, P(C_L \mid x) \]

\[ P(C_1), P(C_2), \ldots, P(C_L) \]
Review: Bayes’ Rule
– for Generative Bayes Classifiers

\[
P(C | X) = \frac{P(X | C)P(C)}{P(X)}
\]

\[P(C_1 | x), P(C_2 | x), \ldots, P(C_L | x)\]

\[P(C / X) = \frac{P(X / C)P(C)}{P(X)}\]
Establishing a probabilistic model for classification through generative modeling

$$\text{argmax}_{c_i} P(C_i | X) = \text{argmax}_{c_i} P(X, C_i) = \text{argmax}_{c_i} P(X | C_i) P(C_i)$$

$P(x | c_1)$

$P(x | c_2)$

$P(x | c_L)$

Generative Probabilistic Model for Class 1

Generative Probabilistic Model for Class 2

Generative Probabilistic Model for Class $L$

$x = (x_1, x_2, \cdots, x_p)$

Adapt from Prof. Ke Chen NB slides
Summary:
Generative classification with the MAP rule

- MAP classification rule
  - **MAP**: Maximum A Posterior
  - Assign $x$ to $c^*$ if

\[
P(C = c^* \mid X = x) > P(C = c \mid X = x) \quad c \neq c^*, \ c = c_1, \ldots, c_L
\]
Summary:
Generative classification with the MAP rule

\[ P(C = c^* \mid X = x) > P(C = c \mid X = x) \quad c \neq c^*, \ c = c_1, \ldots, c_L \]

- Generative classification with the MAP rule
  - Apply Bayes rule to convert them into posterior probabilities

\[
P(C = c_i \mid X = x) = \frac{P(X = x \mid C = c_i)P(C = c_i)}{P(X = x)} \propto P(X = x \mid C = c_i)P(C = c_i) \quad \text{for } i = 1, 2, \ldots, L
\]

- Then apply the MAP rule
**An Example**

- **Example: Play Tennis**

*PlayTennis: training examples*

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
An Example

**Example: Play Tennis**

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

\[ x_1 = \{ \text{sunny, overcast, cool} \}, \quad k_1 = 3 \]

\[ x_2 = \{ \text{hot, mild, cool} \}, \quad k_2 = 3 \]

\[ x_3 = \{ \text{high, normal} \}, \quad k_3 = 2 \]

\[ x_4 = (\text{W, S}), \quad k_4 = 2 \]

\[ C_1 = \{ \text{Yes, No} \}, \quad l = 2 \]

\[ C = \{ \{ \text{Yes}, \text{No} \} \} \]

\[ k_2 = 3 \]

\[ x_2 = \{ \text{Hot, Mild, Cool} \} \]
**Example:** Play Tennis

**PlayTennis:** training examples

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

\[ P(\text{C=Yes}) = \frac{9}{14} \]

\[ P(\text{C=No}) = \frac{5}{14} \]

\[ 3 \times 3 \times 2 \times 2 = 36 \]
\[ P(C = \text{Yes} \mid X_1, X_2, X_3, X_4) \]
\[ P(C = \text{No} \mid X_1, X_2, X_3, X_4) \]

\[ \rightarrow P(C = \text{Yes}) = \frac{9}{14} \]
\[ P(C = \text{No}) = \frac{5}{14} \]

\[ \rightarrow P(X_1, X_2, X_3, X_4 \mid C_i) = \frac{3 \times 3 \times 3 \times 2 \times 2 \times 2}{72} \Rightarrow \text{from train} \]

\[ \text{argmax}_{i=1,2} P(\bar{X}_{ts} \mid C_i) P(C_i) \quad \text{Generative BC} \]
- maximum likelihood estimates
  – simply use the frequencies in the data

\[ p(\text{overcast, hot, high, weak} | \text{Yes}) = \frac{1}{9} \]

### PlayTennis: training examples

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Generative Bayes Classifier:

- Learning Phase

\[ P(C_1), P(C_2), \ldots, P(C_L) \]

\[ P(Play=Yes) = 9/14 \quad P(Play=No) = 5/14 \]

\[ P(X_1, X_2, \ldots, X_p | C_1), P(X_1, X_2, \ldots, X_p | C_2) \]

<table>
<thead>
<tr>
<th>Outlook (3 values)</th>
<th>Temperature (3 values)</th>
<th>Humidity (2 values)</th>
<th>Wind (2 values)</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>weak</td>
<td>0/9</td>
<td>1/5</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>strong</td>
<td>\ldots/9</td>
<td>\ldots/5</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>normal</td>
<td>weak</td>
<td>\ldots/9</td>
<td>\ldots/5</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>normal</td>
<td>strong</td>
<td>\ldots/9</td>
<td>\ldots/5</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>

3*3*2*2 [conjunctions of attributes] * 2 [two classes] = 72 parameters
Generative Bayes Classifier:

\[
[\hat{P}(a_1' | c^*) \cdots \hat{P}(a_p' | c^*)] \hat{P}(c^*) > [\hat{P}(a_1' | c) \cdots \hat{P}(a_p' | c)] \hat{P}(c)
\]

- **Test Phase**
  - Given an unknown instance \( X_{ts} = (a_1', \ldots, a_p') \)
  - **Look up tables** to assign the label \( c^* \) to \( X_{ts} \) if
    \[
    \hat{P}(a_1', \ldots, a_p' | c^*) \hat{P}(c^*) > \hat{P}(a_1', \ldots, a_p' | c) \hat{P}(c),
    \]
    \[c \neq c^*, \ c = c_1, \ldots, c_L\]
  - Given a new instance, \( x' = \) (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)
    \[
    \left\{ \frac{P(x' | Yes)}{P(x' | No)} \frac{P(C=Yes)}{P(C=No)} \right\} = \arg \max \ c \Rightarrow \text{predicted} \ C^*
    \]
Today: Generative Bayes Classifiers

✓ Bayes Classifier
  ▪ MAP classification rule
  ▪ Generative Bayes Classifier
✓ Naïve Bayes Classifier
Naïve Bayes Classifier

• Bayes classification

\[
\arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)
\]

Difficulty: learning the joint probability

• Naïve Bayes classification
  – Assumption that all input attributes are conditionally independent!
Naïve Bayes Classifier

- Bayes classification

$$\arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_p | c_j) P(c_j)$$

Difficulty: learning the joint probability

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!
Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!

\[
P(X_1, X_2, \ldots, X_p | C) = P(X_1 | X_2, \ldots, X_p, C)P(X_2, \ldots, X_p | C)
\]

\[
= P(X_1 | C)P(X_2, \ldots, X_p | C)
\]

\[
= P(X_1 | C)P(X_2 | C) \cdots P(X_p | C)
\]
Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!
    \[
P(X_1, X_2, \ldots, X_p | C) = P(X_1 | C) P(X_2 | C) \cdots P(X_p | C)
    \]
  - MAP classification rule: for a sample \( x = (x_1, x_2, \ldots, x_p) \)
    \[
    [P(x_1 | c^*) \cdots P(x_p | c^*)] P(c^*) > [P(x_1 | c) \cdots P(x_p | c)] P(c),
    \]
    \( c \neq c^*, \quad c = c_1, \ldots, c_L \)
Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!

\[ P(X_1, X_2, \ldots, X_p | C) = P(X_1 | C)P(X_2 | C) \cdots P(X_p | C) \]

- MAP classification rule: for a sample \( x = (x_1, x_2, \ldots, x_p) \)

\[ [P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_p | c)]P(c), \text{ if } c \neq c^* \]

\( c \in \{c_1, \ldots, c_L\} \)

\[ \Rightarrow \text{argmax}_{i=1}^{L} P(c_i)P(x_1 | c_i)P(x_2 | c_2) \cdots P(x_p | c_i) \]
Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!
  \[
P(X_1, X_2, \ldots, X_p | C) = P(X_1 | C)P(X_2 | C) \cdots P(X_p | C)
\]
  - MAP classification rule: for a sample \( x = (x_1, x_2, \ldots, x_p) \)
  \[
  [P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_p | c)]P(c),
  \]
  \( c \neq c^*, \; c = c_1, \ldots, c_L \)

\( \{ s = 1, 2, \ldots, p \} \quad \{ f = 1, 2, \ldots, L \} \)

\( P(X_j | c_i) \)
Naïve Bayes Classifier (for discrete input attributes) - training

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set $S$,

For each target value of $c_i$ ($c_i = c_1, \ldots, c_L$)

$$\hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } S;$$
Naïve Bayes Classifier (for discrete input attributes) - training

- Naïve Bayes Algorithm (for discrete input attributes)

  **Learning Phase**: Given a training set \( S \),

  For each target value of \( c_i \) (\( c_i = c_1, \ldots, c_L \))
  \[
  \hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } S;
  \]

  For every attribute value \( x_{jk} \) of each attribute \( X_j \) (\( j = 1, \ldots, p; \ k = 1, \ldots, K_j \))
  \[
  \hat{P}(X_j = x_{jk} | C = c_i) \leftarrow \text{estimate } P(X_j = x_{jk} | C = c_i) \text{ with examples in } S;
  \]

  **Output**: conditional probability tables; for \( X_j, K_j \times L \) elements
Naïve Bayes Classifier
(for discrete input attributes) - training

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set $S$,
    
    For each target value of $c_i$ ($c_i = c_1, \ldots, c_L$)
    \[
    \hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } S;
    \]
    
    For every attribute value $x_{jk}$ of each attribute $X_j$ ($j = 1, \ldots, p; k = 1, \ldots, K_j$)
    \[
    \hat{P}(X_j = x_{jk} | C = c_i) \leftarrow \text{estimate } P(X_j = x_{jk} | C = c_i) \text{ with examples in } S;
    \]
    
    Output: conditional probability tables; for $X_j, K_j \times L$ elements.
Naïve Bayes
(for discrete input attributes) - testing

- Naïve Bayes Algorithm (for discrete input attributes)
  
  **Test Phase**: Given an unknown instance \( X' = (a'_1, \cdots, a'_p) \)

  Look up tables to assign the label \( c^* \) to \( X' \) if

  \[
  [\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)] \hat{P}(c),
  \]

  \( c \neq c^* \), \( c = c_1, \cdots, c_L \)

  \[
  \prod_{i=1}^{L} P(X'(c_i), P(c_i)) P(c_i) = \prod_{i=1}^{L} P(a'_i | c_i) P(c_i)
  \]
Today: Generative Bayes Classifiers

- Bayes Classifier
  - MAP classification rule
  - Generative Bayes Classifier
- Naïve Bayes Classifier
- NBC for discrete input variables
An Example

- Example: Play Tennis

**PlayTennis: training examples**

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
## An Example

### Example: Play Tennis

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

$k_2 = 3$

$k_3 = 2$

$k_4 = (W, S)$

$C = \{\text{Yes, No}\}$

$X_1 = \{\text{sunny, overcast}\}$

$X_2 = \{\text{hot, mild, cool}\}$

$X_3 = \{\text{high, normal}\}$

$X_4 = \{\text{weak, strong}\}$

PlayTennis: training examples
Learning (training) the NBC Model

\[ P(C_1, \ldots, C_L) \]

\[ \mathcal{L} \]

- \( X_1 \)
- \( X_2 \)
- \( X_3 \)
- \( X_4 \)
- \( X_5 \)
- \( X_6 \)
Learning (training) the NBC Model

- maximum likelihood estimates:
  - simply use the frequencies in the data

\[
\hat{P}(c_j) = \frac{N(C = c_j)}{N}
\]

\[
\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}
\]
Learning (training) the NBC Model

- Maximum likelihood estimates
  - Simply use the frequencies in the data

\[
\hat{P}(c_j) = \frac{N(C = c_j)}{N}
\]

\[
\hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}
\]
\[ C = c_i \]

2 - Dimensional

\[ X_j = x_{jk} \]
## PlayTennis: training examples

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

\[
\Pr(X_1 = \text{Rain} \mid C = \text{Yes}) = \frac{3}{9}
\]

\[
\Pr(X_1 = \text{Rain} \mid C = \text{No}) = \frac{2}{5}
\]
Learning Phase

Estimate $P(X_j = x_{jk} | C = c_i)$ with examples in training:

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>2/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Overcast</td>
<td>4/9</td>
<td>0/5</td>
</tr>
<tr>
<td>Rain</td>
<td>3/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot</td>
<td>2/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Mild</td>
<td>4/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Cool</td>
<td>3/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Humidity</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>3/9</td>
<td>4/5</td>
</tr>
<tr>
<td>Normal</td>
<td>6/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wind</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>3/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Weak</td>
<td>6/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

$P(X_2|C_1), P(X_2|C_2)$

$P(X_4|C_1), P(X_4|C_2)$

$P(Play=Yes) = 9/14 \quad P(Play=No) = 5/14 \quad P(C_1), P(C_2), \ldots, P(C_L)$

3+3+2+2 [naïve assumption] * 2 [two classes]= 20 parameters
Learning Phase

Estimate $P(X_j = x_{jk} | C = c_i)$ with examples in training:

$P(X_2|C_1), P(X_2|C_2)\quad P(X_4|C_1), P(X_4|C_2)$

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>2/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Overcast</td>
<td>4/9</td>
<td>0/5</td>
</tr>
<tr>
<td>Rain</td>
<td>3/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot</td>
<td>2/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Mild</td>
<td>4/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Cool</td>
<td>3/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Humidity</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>3/9</td>
<td>4/5</td>
</tr>
<tr>
<td>Normal</td>
<td>6/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wind</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>3/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Weak</td>
<td>6/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

3+3+2+2 [naïve assumption] * 2 [two classes] = 20 parameters

$P(\text{Play}=\text{Yes}) = 9/14\quad P(\text{Play}=\text{No}) = 5/14\quad P(C_1), P(C_2), \ldots, P(C_L)$
Testing the NBC Model

• Test Phase
  – Given a new instance,
    \[ x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong}) \]
Testing the NBC Model

\[ \hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*) \hat{P}(c^*) > \left[ \hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c) \right] \hat{P}(c) \]

- **Test Phase**
  - Given a new instance,
  \[ x' = (\text{Outlook}=\text{Sunny}, \ \text{Temperature}=\text{Cool}, \ \text{Humidity}=\text{High}, \ \text{Wind}=\text{Strong}) \]

\[ \frac{9}{14} \times \frac{2}{9} \cdots \cdot \cdot = \]

\[ \frac{5}{14} \times \frac{3}{5} \times \cdot \cdot = \]
Testing the NBC Model

• Test Phase
  - Given a new instance,
    \( x' = (\text{Outlook=} \text{Sunny}, \text{Temperature=} \text{Cool}, \text{Humidity=} \text{High}, \text{Wind=} \text{Strong}) \)
  - Look up in conditional-prob tables

\[
\begin{align*}
P(\text{Outlook=} \text{Sunny} \mid \text{Play=} \text{Yes}) &= 2/9 \\
P(\text{Temperature=} \text{Cool} \mid \text{Play=} \text{Yes}) &= 3/9 \\
P(\text{Humidity=} \text{High} \mid \text{Play=} \text{Yes}) &= 3/9 \\
P(\text{Wind=} \text{Strong} \mid \text{Play=} \text{Yes}) &= 3/9 \\
P(\text{Play=} \text{Yes}) &= 9/14 \\
P(\text{Outlook=} \text{Sunny} \mid \text{Play=} \text{No}) &= 3/5 \\
P(\text{Temperature=} \text{Cool} \mid \text{Play=} \text{No}) &= 1/5 \\
P(\text{Humidity=} \text{High} \mid \text{Play=} \text{No}) &= 4/5 \\
P(\text{Wind=} \text{Strong} \mid \text{Play=} \text{No}) &= 3/5 \\
P(\text{Play=} \text{No}) &= 5/14
\end{align*}
\]
Testing the NBC Model

• Test Phase

– Given a new instance, 
\( x' = \{\text{Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong}\} \)

– Look up in conditional-prob tables

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Outlook=Sunny</td>
<td>Play=Yes)</td>
</tr>
<tr>
<td>P(Temperature=Cool</td>
<td>Play=Yes)</td>
</tr>
<tr>
<td>P(Humidity=High</td>
<td>Play=Yes)</td>
</tr>
<tr>
<td>P(Wind=Strong</td>
<td>Play=Yes)</td>
</tr>
<tr>
<td>P(Play=Yes)</td>
<td>9/14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Outlook=Sunny</td>
<td>Play=No)</td>
</tr>
<tr>
<td>P(Temperature=Cool</td>
<td>Play=No)</td>
</tr>
<tr>
<td>P(Humidity=High</td>
<td>Play=No)</td>
</tr>
<tr>
<td>P(Wind=Strong</td>
<td>Play=No)</td>
</tr>
<tr>
<td>P(Play=No)</td>
<td>5/14</td>
</tr>
</tbody>
</table>

– MAP rule

\[ P(\text{Yes} | x') = \frac{P(\text{Sunny} | \text{Yes})P(\text{Cool} | \text{Yes})P(\text{High} | \text{Yes})P(\text{Strong} | \text{Yes})}{P(\text{Play}=\text{Yes})} = 0.0053 \]

\[ P(\text{No} | x') = \frac{P(\text{Sunny} | \text{No})P(\text{Cool} | \text{No})P(\text{High} | \text{No})P(\text{Strong} | \text{No})}{P(\text{Play}=\text{No})} = 0.0206 \]

Given the fact \( P(\text{Yes} | x') < P(\text{No} | x') \), we label \( x' \) to be “No”. 4/5/18
WHY ? Naïve Bayes Assumption

- **$P(c_j)$**
  - Can be estimated from the frequency of classes in the training examples.

- **$P(x_1,x_2,...,x_p | c_j)$**
  - $O(|X_1| \cdot |X_2| \cdot |X_3| \cdot ... \cdot |X_p| \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

If no naïve assumption
WHY ? Naïve Bayes Assumption

• \( P(c_j) \)
  - Can be estimated from the frequency of classes in the training examples.

• \( P(x_1, x_2, ..., x_p | c_j) \)
  - \( O(|X_1| \cdot |X_2| \cdot |X_3| ... \cdot |X_p| \cdot |C|) \) parameters
  - Could only be estimated if a very, very large number of training examples was available.

• \( P(x_k | c_j) \)
  - \( O([|X_1| + |X_2| + |X_3| + ... + |X_p|] \cdot |C|) \) parameters
  - Assume that the probability of observing the conjunct of attributes is equal to the product of the individual probabilities \( P(x_i | c_j) \).
WHY ? Naïve Bayes Assumption

• $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.

• $P(x_1, x_2, ..., x_p | c_j)$
  - $O(|X_1| \cdot |X_2| \cdot |X_3| \cdot \ldots \cdot |X_p| \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

• $P(x_k | c_j)$
  - $O([|X_1| + |X_2| + |X_3| + \ldots + |X_p|] \cdot |C|)$ parameters
  - Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i | c_j)$.
Learning (training) the NBC Model

\[ P(C_1, \ldots, C_k) \]

For instance:

\[ C=\text{Flu} \]

\[ X_6=\text{Muscle-ache} \]
• What if we have seen no training cases where patient had no flu and muscle aches?

• Zero probabilities cannot be conditioned away, no matter the other evidence!

\[
\hat{P}(X_6 = t \mid C = \text{not flu}) = \frac{N(X_6 = t, C = \text{nf})}{N(C = \text{nf})} = 0
\]

\[
?? = \arg \max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)
\]
\[ \delta_f = p(c = \text{flu}) \ p(x_1 | f) \ p(x_2 | f) \ p(x_3 | f) \ p(x_4 | f) \ p(x_5 | f) \ p(x_6 | f) \]

\[ \delta_{nf} = p(c = \text{nf}) \ p(x_1 | nf) \ p(x_2 | nf) \ p(x_3 | nf) \ p(x_4 | nf) \ p(x_5 | nf) \ p(x_6 | nf) \]

if any term gives 0,
\[ \Rightarrow \delta_{nf} = 0 \]
no matter other terms' value
Smoothing to Avoid Overfitting

\[ \hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i} \]

# of values of feature \( X_i \)

To make \( \sum_i (P(x_i | C) = 1) \)

\( \|X_i\| = k_i \)

4/5/18

Adapt From Manning’ textCat tutorial
Smoothing to Avoid Overfitting

\[ \hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i} \]

- Somewhat more subtle version

\[ \hat{P}(x_{i,k} | c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + m p_{i,k}}{N(C = c_j) + m} \]

overall fraction in data where \( X_i = x_{i,k} \)

extent of “smoothing”
Summary:

Generative Bayes Classifier with the MAP rule

Task: Classify a new instance \( X \) based on a tuple of attribute values \( X = \langle X_1, X_2, \ldots, X_p \rangle \) into one of the classes

\[
c_{MAP} = \arg\max_{c_j \in C} P(c_j | x_1, x_2, \ldots, x_p)
\]

\[
= \arg\max_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_p | c_j)P(c_j)}{P(x_1, x_2, \ldots, x_p)}
\]

\[
= \arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_p | c_j)P(c_j)
\]

MAP = Maximum A Posteriori

Adapt From Carols’ prob tutorial
Generative Bayes Classifier

\[
\arg\max_k P(C_k | X) = \arg\max_k P(X, C) = \arg\max_k P(X | C)P(C)
\]

**Task**
- **Classification**
- **EPE with 0-1 loss**
- **Likelihood**
- **Many options**
- **Prob. Models’ Parameters**

**Representation**
- **Score Function**
- **Search/Optimization**
- **Models, Parameters**

**Bernoulli Naïve**

\[
p(W_i = true | c_k) = p_{i,k}
\]

**Gaussian Naïve**

\[
\hat{P}(X_j | C = c_k) = \frac{1}{\sqrt{2\pi} \sigma_{jk}} \exp \left( -\frac{(X_j - \mu_{jk})^2}{2\sigma_{jk}^2} \right)
\]

**Multinomial**

\[
P(W_1 = n_1, \ldots, W_v = n_v | c_k) = \frac{N!}{n_{1k}! n_{2k}! \ldots n_{vk}!} \theta_{1k}^{n_{1k}} \theta_{2k}^{n_{2k}} \ldots \theta_{vk}^{n_{vk}}
\]
References

- Prof. Andrew Moore’s review tutorial
- Prof. Ke Chen NB slides
- Prof. Carlos Guestrin recitation slides